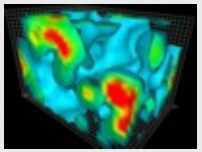
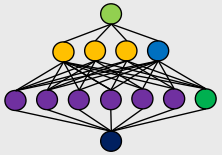
Data Provenance

Introduction and Survey





Luciano Piccoli

IIT/Fermilab

September 23 2008

Outline

- Provenance
  - Definitions
  - e-Science and Workflows
  - Taxonomy
- Provenance Systems
  - Components
  - Systems
- LQCD Provenance

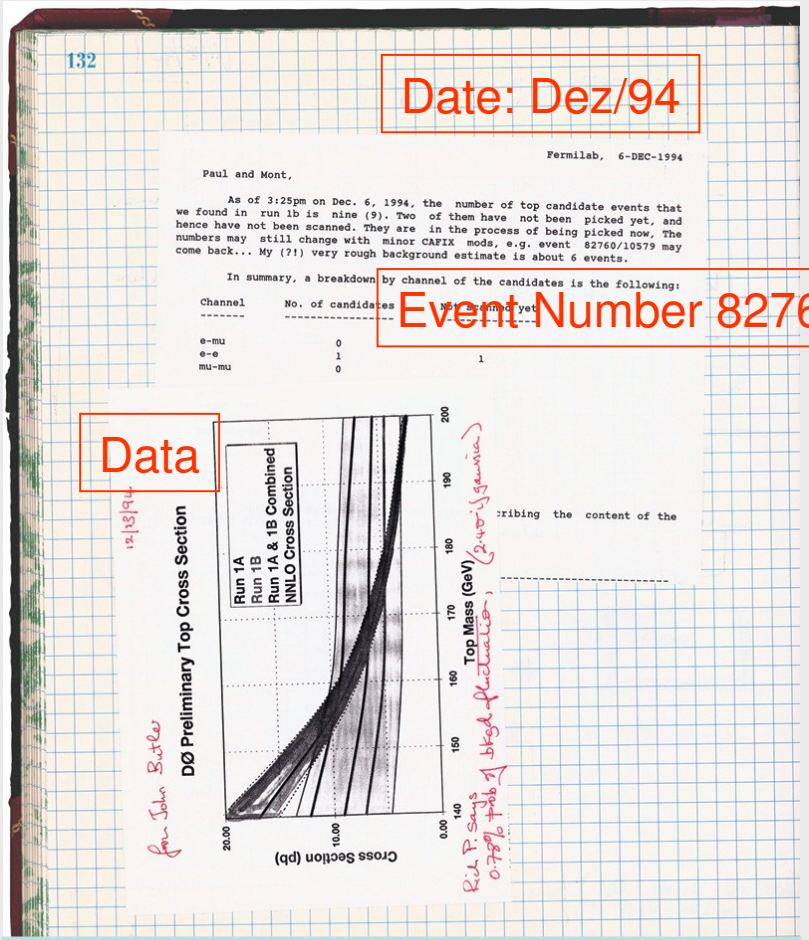
Provenance Definitions

- Oxford English Dictionary: (i) the fact of coming from some particular **source** or quarter; **origin**, **derivation**. (ii) the **history** or **pedigree** of a work of art, manuscript, rare book, etc.; concretely, a record of the ultimate derivation and passage of an item through its various owners.
- Wikipedia: from the French provenir, "to come from", means the **origin**, or the **source**, of something, or the **history** of the **ownership** or **location** of an object, especially a work of art, or some object of value such as is found in archaeology, or paleontology, or some document, such as a manuscript, or even an item of literature in the broadest sense, including a first edition of a very rare published work. The primary purpose of provenance is to confirm the **time**, **place**, and if appropriate the **person responsible**, for the creation, production or discovery of the object. Comparative techniques, expert opinions, written and verbal records and the results of various kinds of scientific tests are often used to help establish provenance.

source: http://www.cs.utah.edu/~juliana/pub/SIGMOD2008/sigmod2008-tutorial-final.pdf and wikipedia data provenance

Provenance & Science

- Not a new issue
- Lab notebooks have been used for a long time
  - Reproduce results
  - Evidence in patent disputes
- What is new?
  - Large volumes of data
  - Complex analyses
- Writing notes is no longer an option
  - Need systematic means to capture provenance



Logbook entry showing Top Quark discovery at Fermilab

partial source: http://www.cs.utah.edu/~juliana/pub/SIGMOD2008/sigmod2008-tutorial-final.pdf

e-Science

- The term e-Science is used to describe computationally intensive science that is carried out in highly distributed network environments, or science that uses immense data sets that require grid computing;
- Examples e-Science
  - Particle physics
  - Earth sciences
  - Bio-informatics
- Particle physics has a particularly well developed e-Science infrastructure due to their need for adequate computing facilities for the analysis of results and storage of data originating from the CERN Large Hadron Collider.

source: wikipedia e-science

Scientific Workflows

- Special type of workflow that often underlies many large-scale complex e-science applications such as climate modeling, structural biology and chemistry, medical surgery or disaster recovery simulation.
- Compared with business workflows, scientific workflow has special features such as computation, data or transaction intensity, less human interaction, and a large number of activities.
- Help domain scientists to focus on the science instead of spending time writing scripts and monitoring submitted computations.

Provenance & Scientific Workflows

- Scientific workflows carry out work for e-Science experiments.
- Workflow systems can save execution information by tracing input and output data generated (no more logbooks).
- Data provenance is a secondary product of running a scientific workflow, but many systems do not record it!

Why Provenance?

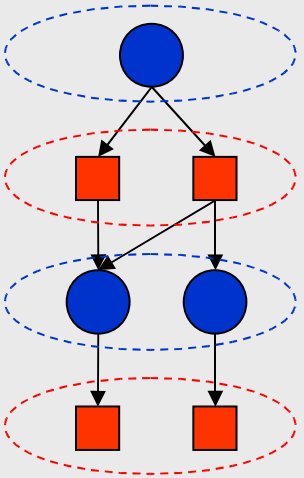
- Understanding of data and processes
- Quality assurance: validation, repudiation and debugging
- Data Quality: estimate quality and reliability of data based on the source data and transformations recorded by the provenance. Provides proof statements of derivation
- Audit Trail: resource usage, detect errors in data generation and generation trace for audits
- Replication Recipes: allow repetition of data derivation
- Attribution: help establish copyright and ownership of data
- Informational: provide a context to interpret data

partial source: Y. N. Simmhan, B. Plale and D. Gannon, A Survey of Data Provenance in e-Science. SIGMOD Record, Vol. 34, No. 3, Sept. 2005.

Taxonomy

Subject of Provenance

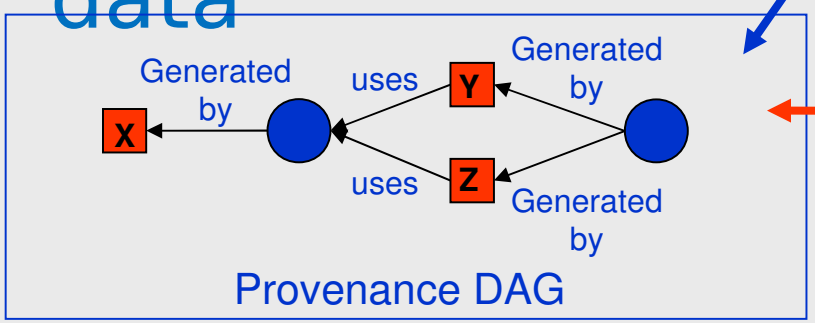
- Data vs. Process
  - Data-oriented: provenance gathered from the data products
  - Process-oriented: provenance captured through examination of input and output data
- Granularity
  - Amount vs. usefulness of data
  - Storage cost inversely proportional to granularity



source: Y. N. Simmhan, B. Plale and D. Gannon, A Survey of Data Provenance in e-Science. SIGMOD Record, Vol. 34, No. 3, Sept. 2005.

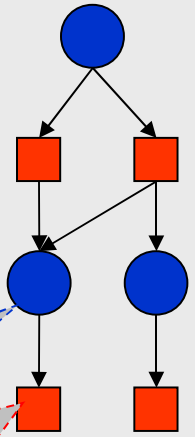
Representation

- Annotations: descriptions about source data and processes
- Inversion: derivations are inverted to find the input data data



File X = Process A (1, 2, 3, File Y, File Z)

File X, produced by process A, using parameters 1, 2 and 3, and files Y and Z



partial source: Y. N. Simmhan, B. Plale and D. Gannon, A Survey of Data Provenance in e-Science. SIGMOD Record, Vol. 34, No. 3, Sept. 2005.

Representation

- No language standard so far
- XML for representing lineage information
- Semantic Web technologies
  - RDF – Resource Description Framework (http://www.w3schools.com/RDF/default.asp)
  - OWL – Web Ontology Language (http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html)
- Open Provenance Model v1.01 (http://eprints.ecs.soton.ac.uk/16148/1/opm-v1.01.pdf)
  - Recent – July 2008
  - Very generic

partial source: Y. N. Simmhan, B. Plale and D. Gannon, A Survey of Data Provenance in e-Science. SIGMOD Record, Vol. 34, No. 3, Sept. 2005.

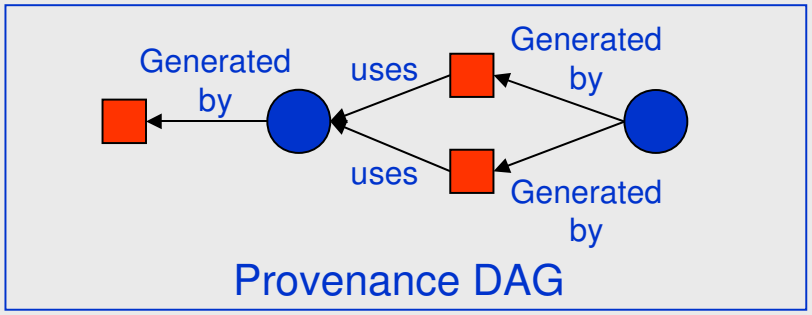
Storage

- Granularity
  - Provenance data may be larger than the scientific data!
  - Tuples in database
  - Collections of provenance log files
- Scalability
  - Inversion (usually less data) more scalable than annotation
- Where?
  - Embedded within data file
    - Provenance header
    - Easy to maintain integrity of provenance
    - Hard to make searches
  - Separate file or database
    - Synchronize metadata and data
    - Collection mechanisms
    - Extra costs for storage and collection

source: Y. N. Simmhan, B. Plale and D. Gannon, A Survey of Data Provenance in e-Science. SIGMOD Record, Vol. 34, No. 3, Sept. 2005.

Access

- Derivation graphs

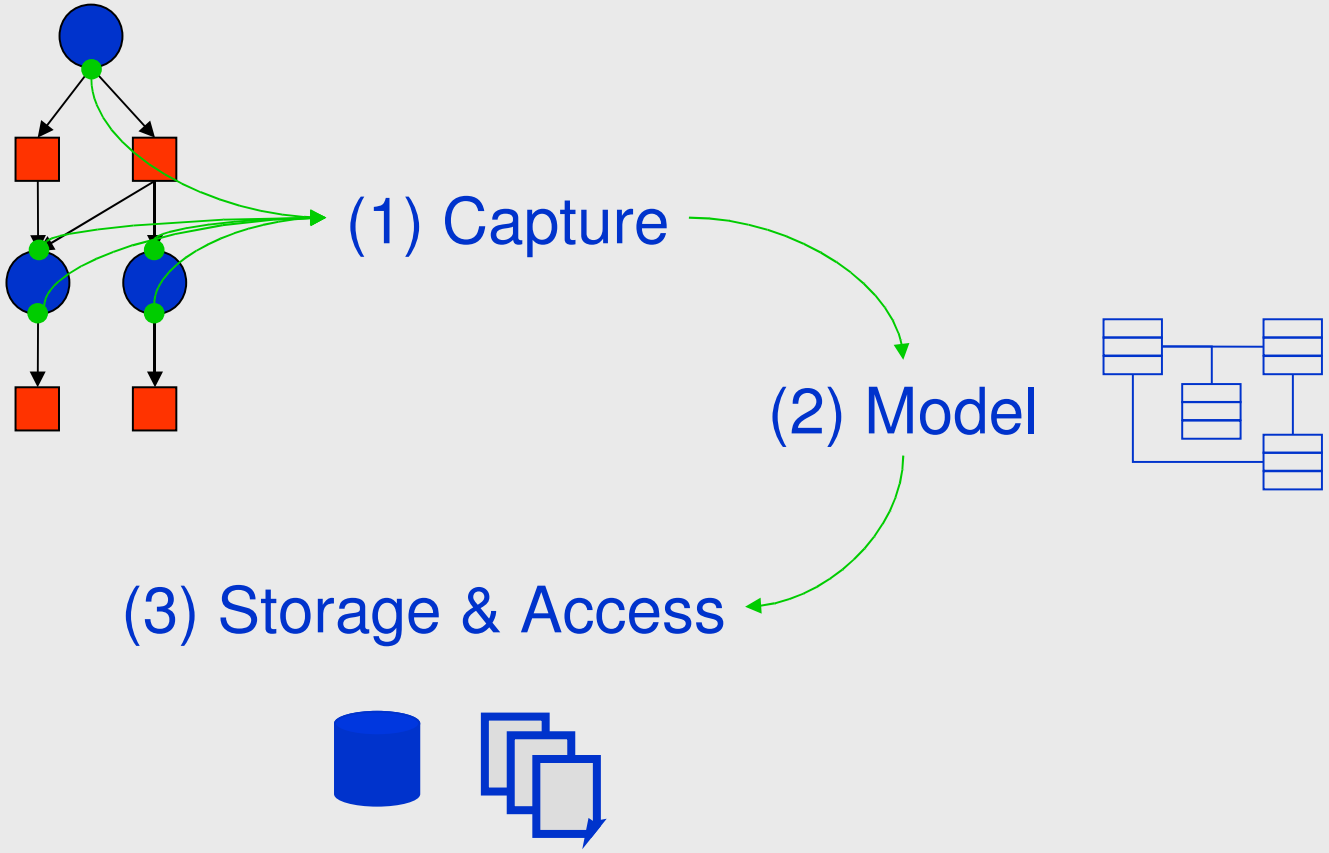


- Search for datasets on the provenance data
  - Use results as input to workflows
  - Search languages (e.g. SQL) not designed for provenance

partial sources: Y. N. Simmhan, B. Plale and D. Gannon, A Survey of Data Provenance in e-Science. SIGMOD Record, Vol. 34, No. 3, Sept. 2005 and J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for computational tasks: A survey. Computing in Science and Engineering, 10(3):11–21, 2008.

Provenance Systems

Components



Components

- Capture Mechanisms
  - Access to job details
    - Execution information
    - User-specified annotations
  - Types
    - Workflow-based
      - Tight coupled with workflow systems
      - Enables straightforward capture process
      - Workflow systems extended to support provenance, same are now designed with provenance enabled
    - Process-based
      - Instrumentation of process to capture provenance
    - OS-based
      - Decoupled from workflows and processes
      - Post-processing for defining relationships between system calls and tasks

source: J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for computational tasks: A survey. Computing in Science and Engineering, 10(3):11–21, 2008.

Components

- Provenance Models
  - Tend to vary according to domain and user needs.
  - What information to store?
    - Process ant data dependencies
    - Scheduling and execution information
  - Workflow systems like Taverna and VisTrails provide provenance models, while others, like Kepler, were extended to support provenance.
- Storage Access
  - Several models: XML, RDF, OWL, SQL
  - File svstems or relational databases

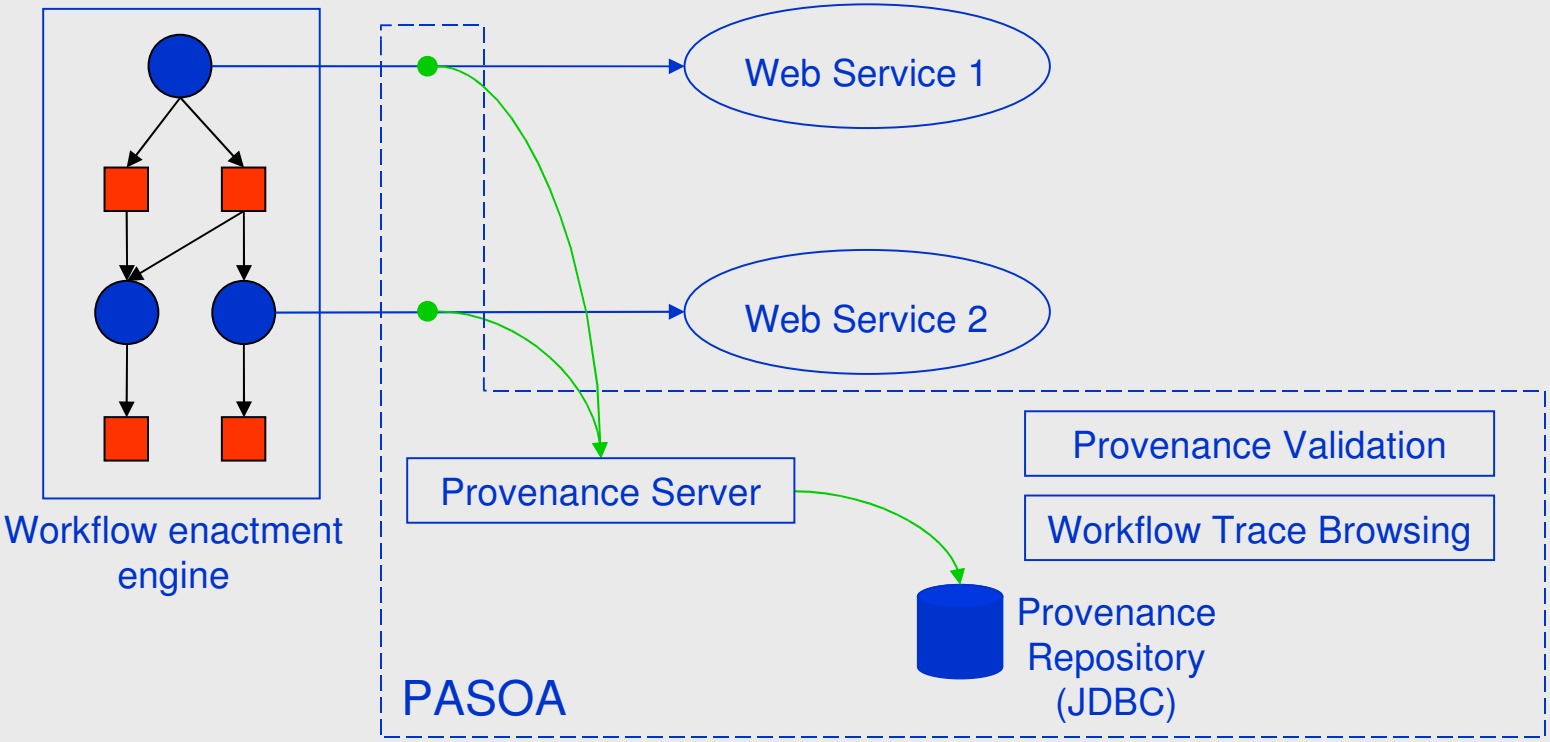
source: J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for computational tasks: A survey. Computing in Science and Engineering, 10(3):11–21, 2008.

Provenance Systems

- Non-workflow based
  - PASOA
  - PASS
  - Karma
- Workflow-based
  - PEGASUS
  - KEPLER

PASOA (Southampton)

- Provenance Aware Service-Oriented Architecture



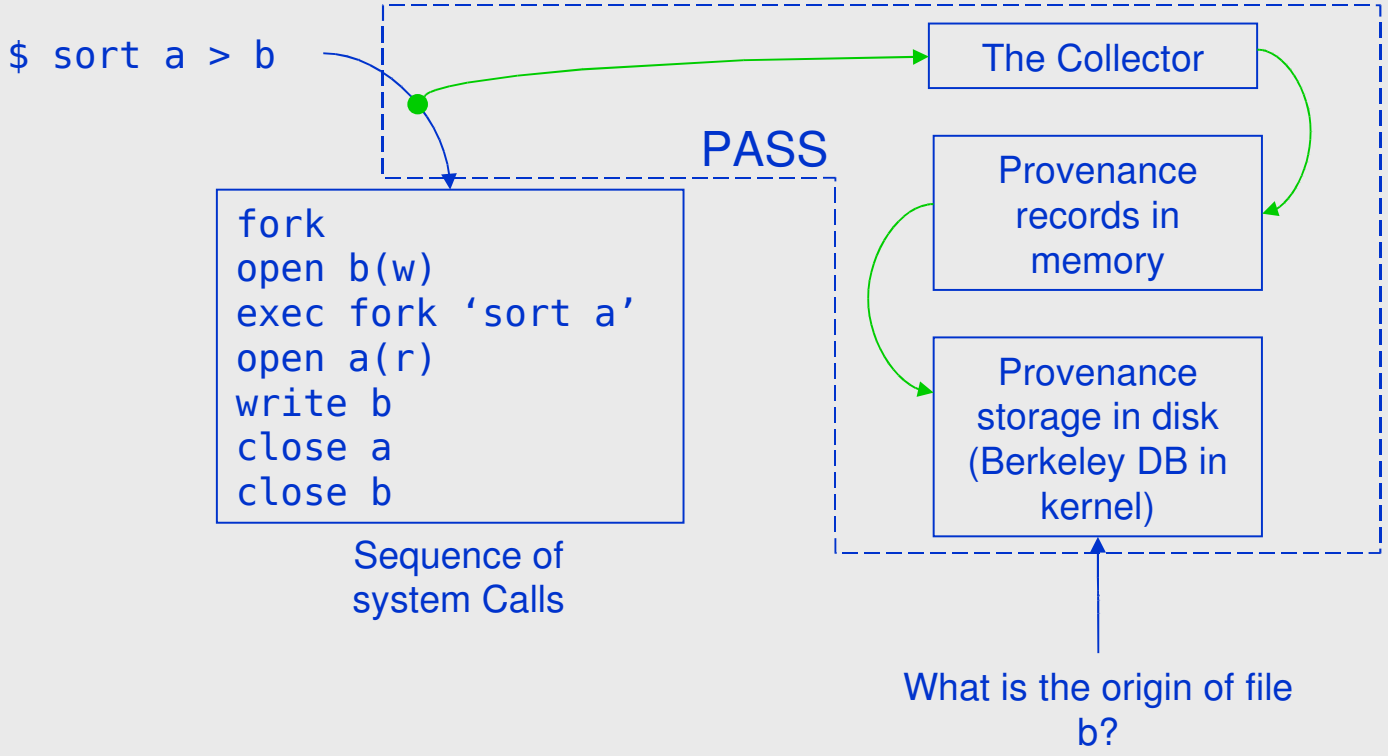
sources: [1] S. Miles, E. Deelman, P. Groth, K. Vahi, G. Mehta, and L. Moreau. Connecting scientific data to scientific experiments with provenance. In E-SCIENCE '07: Proceedings of the Third IEEE International Conference on e-Science and Grid Computing, pages 179–186, Washington, DC, USA, 2007. IEEE Computer Society. [2].J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for computational tasks: A survey. Computing in Science and Engineering, 10(3):11–21, 2008.

PASOA

- Independent of workflow system
- Record workflow traces
- Relational database model
- Trace validation
  - Check if web services still respond the same way
- Trace browsing tool

PASS (Harvard)

- Provenance-Aware Storage Systems
- Operating system level



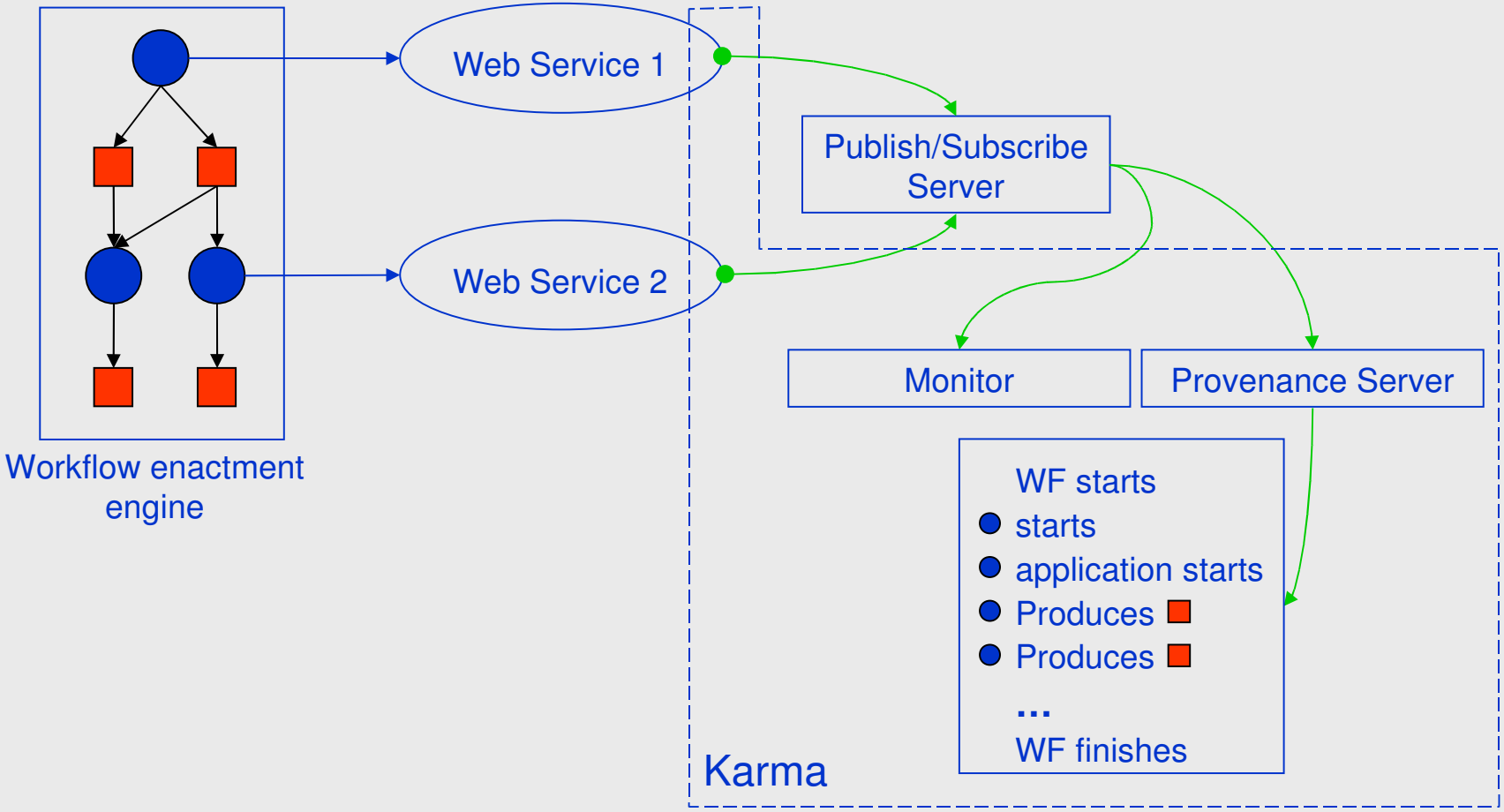
source: K.-K. Muniswamy-Reddy, D. A. Holland, U. Braun, and M. Seltzer. Provenance-aware storage systems. In ATEC '06: Proceedings of the annual conference on USENIX '06 Annual Technical Conference, pages 4–4, Berkeley, CA, USA, 2006. USENIX Association.

PASS

- Completely transparent, no need to instrument code.
- Capture lots of information, even those not needed by the workflow.
- Queries return loads of data, it may be difficult to find what one is looking for.
- Specific to one OS (Linux), no Grid connection.
- Great system admin and debuging tool.

source: J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for computational tasks: A survey. Computing in Science and Engineering, 10(3):11–21, 2008.

Karma (Indiana)



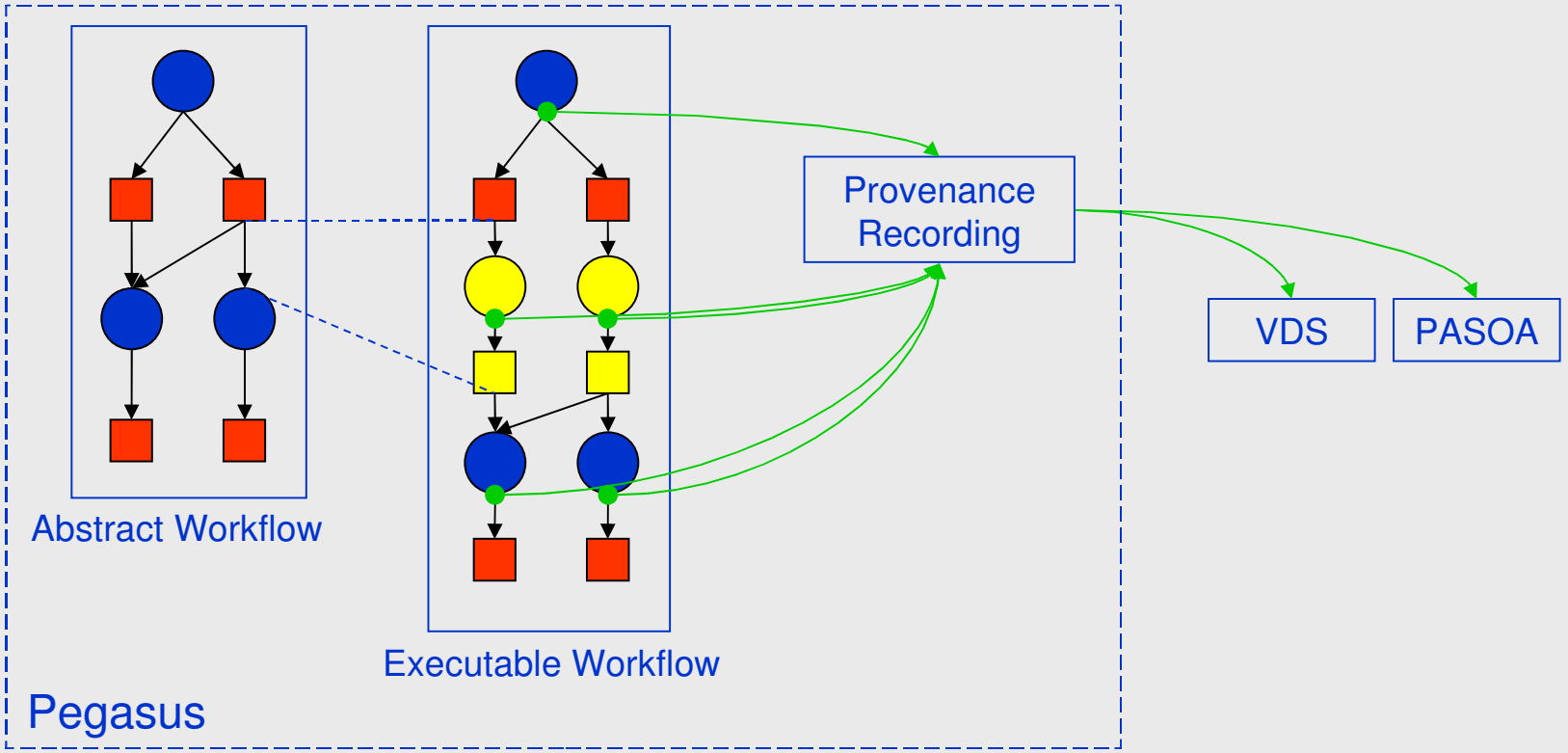
source: Y. L. Simmhan, B. Plale, and D. Gannon. A framework for collecting provenance in data-centric scientific workflows. icws, 0:427–436, 2006.

Karma

- Independent of workflow system
- Dependent on the publish/subscribe mechanism
  - Loosely coupled
  - Uses their own P/S implementation WS-Messenger

Pegasus (USC)

- Workflow Management System and Execution Planner



source: S. Miles, E. Deelman, P. Groth, K. Vahi, G. Mehta, and L. Moreau. Connecting scientific data to scientific experiments with provenance. In E-SCIENCE '07: Proceedings of the Third IEEE International Conference on e-Science and Grid Computing, pages 179–186, Washington, DC, USA, 2007. IEEE Computer Society.

Kepler (SDSC and others)

- Workflow Management System
- Runs the Ptolemy workflow engine
- No explicit provenance support
- Provenance extensions – modification of actors/directors
  - Matrioshka (S. M. S. da Cruz, P. M. Barros, P. M. Bisch, M. L. M. Campos, and M. Mattoso. Provenance services for distributed workflows. In CCGRID, pages 526–533, 2008.)
  - Pipelined workflows (S. Bowers, T. M. McPhillips, B. Lud¨ascher, S. Cohen, and S. B. Davidson. A model for user-oriented data provenance in pipelined scientific workflows. In IPAW, pages 133–147, 2006.)

source: I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludascher, and S. Mock. Kepler: an extensible system for design and execution of scientific workflows. In Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on, pages 423–424, 2004.

LQCD Provenance

LQCD Provenance

- Similar to PASOA
- Capture mechanism independent of Workflow System (Process-based)
  - Mechanism added to participant wrapper instead of web service
- Advantages
  - Flexible data model
  - Query using Ruby language

```
su3 = ParticipantType.find(:first,
                  :conditions => "name == 'su3'")
su3.participant.participant_instances.products
```

- For more information on the design and prototype see previous group meeting slides:
(http://216.47.150.202/wiki/lib/exe/fetch.php?id=group_meeting&cache=cache&media=lucianopiccoli-29-jul-2008.pdf)

LQCD Provenance: Future work

- Workflow Integration
  - Data movement and file management
    - Saving intermediate and final data from worker node – currently files generated on shared area.
  - Participants independent of workflow system
    - Each workflow system requires participants to use different wrappers
  - Integrate database with workflow
    - Import parameters into workflow
  - Remote workflow execution
    - Update main database with remote execution results
- Fault Tolerance
  - Resume from last milestone: resume workflow based on information from the provenance database.
    - Coordinate with workflow system

LQCD Provenance: Future work

- Provenance
  - Data collection
  - Reconstruct provenance from database information
    - Able to trace binaries, input and output files.
    - Provenance graph (e.g. DOT files), including participants and data products
  - Add workflow to provenance schema
    - Dependent of workflow system
    - Database points to external repository (CVS, SVN)
    - Support for multiple workflow languages?
    - Description on how workflow is invoked
      - Automatic creation of web interface where users select versions of participants to use (workflow instantiation)

Thank you